TOWARDS PERCEPTUAL SOUNDSCAPE CHARACTERIZATION USING EVENT DETECTION ALGORITHMS

Félix Gontier[1]*, Pierre Aumond*[3]*, Mathieu Lagrange*[1]*,*
Catherine Lavandier[2]*, Jean-Francois Petiot*[1]

[1] LS2N, UMR 6004, Ecole Centrale de Nantes, CNRS, 44322 Nantes, France, {felix.gontier}@ls2n.fr
[2] ETIS, UMR 8051, Université Paris Seine, Université de Cergy-Pontoise, ENSEA, CNRS, 95000 Cergy-Pontoise, France, {catherine.lavandier}@u-cergy.fr
[3] UMRAE, Ifsttar, 44341 Bouguenais, France, {pierre.aumond}@ifsttar.fr

ABSTRACT

Assessing properties about specific sound sources is important to characterize better the perception of urban sound environments. In order to produce perceptively motivated noise maps, we argue that it is possible to consider the data produced by acoustic sensor networks. In order to validate this important assumption, this paper reports on a perceptual test on simulated sound scenes for which both perceptual and acoustic source properties are known. Results show that it is indeed feasible to predict perceptual source-specific quantities of interest from recordings, leading to the new task of automatic soundscape characterization which specificities are discussed.

***Index Terms*—**

1. INTRODUCTION

The ongoing urbanization process has led to an increase in sound quality concerns. In urban areas the noise has been linked to several health issues including sleep-related troubles as well as heart diseases rates, and is a major cause for city dwellers' annoyance in certain areas. In this context, the 2002/49/CE European directive [1] requires that large cities maintain noise maps to facilitate the development of noise reducing plans. These noise maps are mainly based on predictive maps generated using propagation models. The studies are also 1) often limited to traffic and other transportation sources, and 2) few physical measurements are used. Furthermore, the models depend on topological data that may be unavailable or incomplete. The advent of the Internet of Things (IoT) presents an opportunity for the development of large, scalable networks of acoustic sensors [2, 3]. The Characterization of Urban Sound Environments (CENSE) project [4] aims at implementing such a network to produce perceptually motivated noise maps.

Schafer [5] defined a soundscape as the perceptual result of a sound environment. The assessment of subjective descriptors [6, 7, 8] such as the liveliness or familiarity is thus necessary to evaluate the quality of urban scenes. The relevant attributes describing the appreciation of soundscapes can be mapped in perceptual spaces [9, 10]. The set of considered attributes is reduced to a few dimensions which are used as a basis for perceptual experiments. Specifically, the dimension of pleasantness is increasingly associated with soundscape quality in recent works [11, 12, 13, 14]. Soundscape perception is highly dependent on the composition of the scene [15, 16]. Indeed, each sound source yields a different perceptual response. For example, the soundscape quality is likely to be improved by birdsongs and deteriorated by construction noises.

Acoustic monitoring applications typically rely on the measurement of energetic (sound levels, eg. L_{Aeq}) and psychoacoustic (eg. Zwicker's loudness N) indicators. These global quantities describe efficiently the overall activity, with percentile values linked to event or background assessment. However they do not differentiate sources and are thus not sufficient to a perceptual characterization of soundscapes. Additional information about the typology of active sources and their repartition in time is needed. Several sets of relevant indicators have been studied [17, 18, 19] to better account for the specificities of each scene. Though, the established metrics are mostly derived from the abovementioned physical indicators or are unrelated to sound measurement such as the traffic flow density.

The use of large-scale sensor networks yields a problematic for the extraction of content-related quantities of interest from important amounts of data. Despite a growing interest in the community, machine learning models - to the best of our knowledge - were not yet specifically targeted to the prediction of source-specific perceptual parameters in complex urban environments. Most event detection applications focus on obtaining a precise annotation of source activity, within usual ranges of tens of milliseconds. The estimation of sound levels involves entirely different models through source separation and regression [**?**]. Both detection and level estimation are necessary to predict pleasantness although perceptual notions are evaluated on a coarser time scale and rougher precision than usually considered.

We believe that the use of machine listening techniques could greatly benefit the automatic assessment of urban soundscape quality using sensor networks. Though, as it will be discussed in this paper, some important adaptations needs to be performed to the definition of objectives of event detection systems in order to efficiently tackle this task. The aim of this paper is to 1) bring some context of soundscape pleasantness characterization, and 2) report on a perceptual experiment performed in order to study which features shall be brought by automatic event detection systems in order to gather relevant information for the task of characterize perceptual attributes of the soundscape.

2. SOUNDSCAPE CHARACTERIZATION

Several perceptual studies of the urban soundscape quality have been led to propose a model of pleasantness from other perceptual parameters [20, 9, 14, 13]. In all cases, a good approximation of

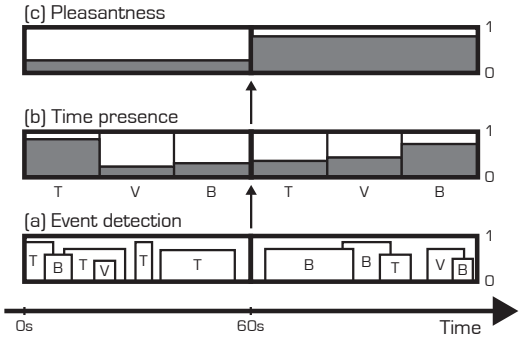

Figure 1: The three suggested levels of metrics to predict soundscape pleasantness. (a) Traffic (T), voice (V) and bird (B) events are detected and their sound level roughly estimated. (b) The perceptual time of presence for each source is computed on one-minute frames, resulting in a pleasantness value (c).

pleasantness can be obtained by linear combination of both overall and source-specific parameters evaluated on discrete scales. Global parameters consider the sound scene in its entirety, the overall loudness is commonly used. The parameters used for the assessment of source-wise contributions include 1) the sound level where each source is considered separately, 2) the emergence or dominance relating to the influence of the source in the global mix, or 3) the time of presence, that is the ratio of time where the source s is heard in a given scene. The notion of time of presence is particularly interesting as it hints at the possibility of automatic prediction through event detection systems. The corresponding model is:

$$P = aL + \sum_s b_s T_s + c \quad (1)$$

where L is the perceived overall level of the scene and T_s is the perceived time of presence for source s. The coefficients a, b_s and c are usually found via multiple linear regression and thus differ in each study. Furthermore, three principal source categories are usually identified: mechanical, human and animal. Mechanical sounds are mainly composed of traffic and are mostly found to have a negative impact on soundscape pleasantness, whereas animal sources (bird activity) have a positive influence and human sounds (voices) can yield mixed effects.

Assuming this perceptual model, the prediction of pleasantness can be assimilated as that of perceived times of presence. These quantities of interest are related to the audio signal and their assessment represents an interesting application for detection models proposed in the data science and machine learning community.

Three levels of metrics are thus identified for this task, as shown in Figure 1. First, the physical level (a) is evaluated on the presence and emergence of the three identified sound sources: traffic (T), voice (V) and birds (B). The needed precision of event onset and sound level estimations is further studied in this paper. The second level (b) is the perceived time of presence for each source in the whole scene represented as a scalar in the 0-1 range. The third level (c) is the estimate of pleasantness over the scene, also represented as a 0-1 scalar. Both the perceptual levels of metrics are only relevant on longer time scales, about one minute being a usual value in existing experiments.

The transition model between the perceived time of presence per source and pleasantness is known. However no previous work exists that uses detection models for the estimation of source-specific subjective parameters. The feasibility of assessing source perception from the postulated metrics at the physical level shall be verified as a first step prior to building the full estimation model.

3. FROM PHYSICAL TO PERCEPTUAL TIME OF PRESENCE OF SOURCES

We thus conduct a perceptual experiment to validate this key step of the estimation chain. Its objectives are to study the relation between extracted source-dependent physical indicators to their perceptual equivalents, then validate the relevance of the first level of metrics introduced in the previous section.

For this perceptual test, a set of sound scenes recorded in the 13th district of Paris as part of the GRAFIC project [14] is used as reference. Some artificial scenes with equivalent event sequencing are also used for which the acoustic properties of each source of the scene can be computed precisely.

Of the 19 different recording locations, 9 are selected to represent diverse compositional properties: park (P3, P9), quiet street (P5, P11, P13, P17), noisy street (P2, P6) and very noisy street (P16). Corresponding artificial scenes are simulated [21] using *simScene* [22]. To do so, the recordings are first annotated by identifying active background and event sources. Background sounds are present throughout the whole scene and are characterized by an absolute level parameter.

Conversely, events are localized occurrences that are defined by their start and end times as well as an event-to-background ratio (EBR). The sound scenes are simulated from these annotations and a database of extracts for isolated sources. This ensures that ground truth source-specific presence and sound level can be computed. One minute of audio is extracted for each scene such as no single event overwhelms the perception of the rest of the excerpt.

During the test, the order of appearance is as follows: the original recorded scenes from locations P3 and P16 representing very quiet (park) and very noisy (very noisy street) environments are always presented first to calibrate the subject's answers. The 9 simulated sounds are then presented in random order to limit order biases over the participants population. For each scene, 14 criteria are evaluated on a 0-10 scale by the subject. The first four questions cover general sound level and perceptual parameters:

1. *Noisy - Quiet*: Overall perceived loudness (OL),
2. *Boring, uninteresting - Stimulating, interesting*: Interest (I),
3. *Inert, amorphous - Lively, eventful*: Liveliness (L),
4. *Agitated, chaotic - Calm, peaceful*: Calmness (C).

Source-specific perceived time of presence (scale *Jamais - Continuellement*) and sound level (scale *Very low - Very high*) are also evaluated. The considered sources are traffic (T), birds (B), horns and sirens (H), human voice (V) and footsteps (F). Perceived time of presence and level for source S are respectively noted S_T and S_L in the remainder of this paper.

Participants can only listen to each scene once and must answer all questions before listening to the next scene. All subjects used the same hardware desktop configuration, sound card and software, as well as Beyerdynamics DT-990 headphones in a quiet environment. The same sound output level was set by the experimenter for all scenes and participants. 30 subjects took the test in 3 sessions, all reported normal hearing conditions.

An outlier detection procedure is applied on the 270 resulting assessments (30 subjects, 9 scenes). An assessment is rejected when

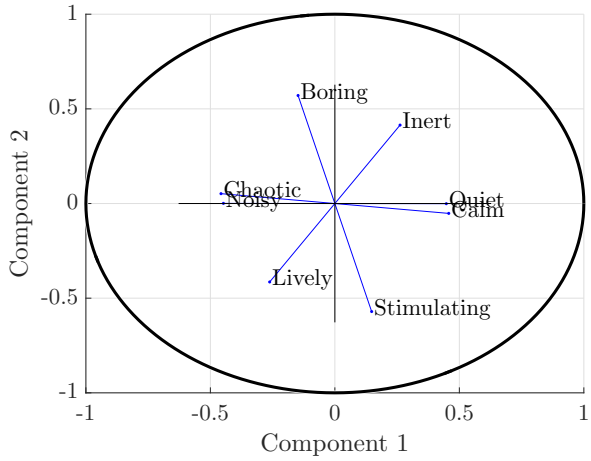

Figure 2: Principal component analysis (first two components) of the four general perceptual parameters at the scene level (n=9). The observed space is distorted although comparable that of previous works in the literature.

it is located away from the mean by more than 3 standard deviations at the question level, that is for each parameter of each scene. The results from two participants with more than 10% assessments considered as outliers are removed from the study, which thus includes 252 individual assessments on 20 parameters.

The perceptual space produced by the test is first compared to previous studies in the literature. This is to ensure that relevant conclusions can be made on further analysis. Figure 2 shows the principal component analysis (PCA) of the four general questions at the scene level (n=9). The first two components respectively explain 52.3% and 30.5% of the global variance. It is found that liveliness (L) correlates poorly with calmness (C), while interest (I) is between the two. These results can be compared to previous studies on similar pleasantness parameters [9, 10, 23], where interest and calmness were established as almost independent. The scale of liveliness was in both cases correlated similarly with the the two others. However, just as the principal components space is slightly distorted in [23] due to the study being focused on park environments, here the small size of the considered corpus in terms of number of scenes also affect the relations between perceptual parameters.

As discussed in Section 2 several models have been established to assess pleasantness as a function of global and source-specific parameters. The main objective of this work is to predict pleasantness from acoustical data without perceptual assessments. Thus, physical indicators are computed from the audio tracks obtained during scene simulation. To evaluate the overall loudness of the scene three measurements are chosen in accordance previous studies [11, 13, 14]:

- L50: Z-weighted (no weighting over the observed frequency range) sound level exceeded 50% of the time in dB,
- LA50: A-weighted sound level exceeded 50% of the time in dBA,
- L50 for the 1kHz band only.

Source-specific indicators are also computed: the time of presence and an emergence estimation metric (resp. T_s and L_s for source s), obtained by subtracting the global L90 (Z-weighted level exceeded 90% of the time) found to represent well background activity to the L10 of each source. The emergence is considered for the whole source activity while time of presence can only be associated to sound events. Sound levels are computed with the Matlab ITA toolbox [24] in the 20 Hz-20 kHz range.

The time-frequency second derivative (TFSD) indicator was shown to improve pleasantness predictions in [14]. Two TFSD indicators are additionally computed: the TFSD at 4 kHz and 125 ms frame duration which is found to strongly correlate with the perceived bird presence, and at 500 Hz and 1 s frame duration which correlates with human voice presence.

In the considered scenes, background sources are always active. The measurement of time of presence is thus limited to sound events which leads to relatively poor representation of the ground truth. This is particularly problematic for traffic sources active in the background of most urban soundscapes. Furthermore, the considered indicators are computed for each sound source separately, not taking into account potential masking effects by other sources active at the same time. Two additional indicators are thus designed regarding these considerations.

The first indicator $T_s(\alpha)$ is a time of presence metric relying on the emergence of each sound source relative to the others over time. Sound levels (dB SPL) are computed for audio frames of 125 ms. This duration approximately corresponds to that of the shortest found event and is widely used in acoustical monitoring applications. The emergence, *i.e.* difference $\Delta_s(t)$ of sound levels between the studied source ($L_s(t)$) and the background constituted of all others ($L_b(t)$) is computed. The source is then considered present on a given time frame if the emergence is greater than a threshold value α. Finally, a time of presence measurement is obtained by averaging over time:

$$T_s(\alpha) = \frac{1}{N_t} \sum_{t=1}^{N_t} \mathbb{1}_{\Delta_s(t)>\alpha} \tag{2}$$

where N_t is the total number of 125 ms analysis frames in the scene. The optimal threshold is found via grid search to be $\alpha = -31dB$ for the considered corpus. If a source is objectively present it will almost always considered present regardless of the other contents of the scene.

However, the masking of a sound by another does not depend only on the emergence over the whole frequency spectrum. The spectral distribution is important. In a signal with a monophonic tone and large band noise, the noise has to be at a higher level to mask completely the tone. The level comparison must be made around the frequency of the tone. A second indicator $T_s(\alpha, \beta)$, based on a spectral decomposition is thus proposed. Third-octave bands sound levels are computed on 125 ms frames and the emergence of a source compared to the background is defined as

$$\Delta_s(t, f) = L_s(t, f) - L_b(t, f) \tag{3}$$

Similarly to the first metric $T_s(\alpha, \beta)$ then relies on simple thresholds applied on the emergence, first in frequency then in time. Its expression is as follows:

$$T_s(\alpha, \beta) = \frac{1}{N_t} \sum_{t=1}^{N_t} \mathbb{1} \left[\frac{\sum_{f=1}^{N_f} \Delta_s(t, f) \mathbb{1}_{\Delta_s(t,f)>\alpha}}{\sum_{f=1}^{N_f} \mathbb{1}_{\Delta_s(t,f)>\alpha}} > \beta \right] \tag{4}$$

where N_f is the number of third-octave bands. Again, optimal values for parameters $\alpha_{opt} = -6dB$ and $\beta_{opt} = -5dB$ are found via

Table 1: Pearson correlation coefficients between perceptual parameters and physical indicators at the scene level (n=9). *: $p < 0.05$, **: $p < 0.01$, non-significant correlations ($p > 0.05$) are noted NS.

Phys./Perc.	OL	I	L	C	T_L	T_T	B_L	B_T	H_L	H_T	V_L	V_T	F_L	F_T
$L50_{1kHz}$	0.93**	NS	NS	-0.92**	0.75*	0.7*	NS	NS	NS	NS	NS	NS	NS	NS
$L50$	0.98**	NS	0.73*	-0.97**	0.72*	NS	NS	NS	NS	NS	NS	NS	NS	NS
$LA50$	0.96**	NS	0.73*	-0.94**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
$TFSD_{500Hz}$	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
$TFSD_{4kHz}$	NS	0.88**	NS	NS	-0.72*	-0.71*	0.92**	0.81**	NS	NS	NS	NS	NS	NS
T_T	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
L_T	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
T_B	NS	0.67*	NS	NS	0.71*	0.75*	NS	NS	NS	NS	NS	NS	NS	NS
L_B	NS	0.93**	NS	NS	-0.84**	-0.83**	0.91**	0.82**	NS	NS	NS	NS	NS	NS
T_H	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.84**	NS	NS	NS	NS
L_H	NS	NS	NS	NS	NS	NS	NS	NS	0.98**	0.78*	NS	NS	NS	NS
T_V	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
L_V	NS	NS	0.81**	NS	NS	NS	NS	NS	NS	NS	0.84**	0.88**	NS	NS
T_F	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.9**	0.68*
L_F	NS	NS	-0.72*	NS	NS	NS	NS	NS	NS	NS	-0.69*	-0.78*	0.92**	NS
$T_T(\alpha)$	NS	-0.81**	NS	NS	0.90**	0.94**	NS	NS	NS	NS	NS	NS	NS	NS
$T_T(\alpha,\beta)$	NS	-0.80**	NS	NS	0.88**	0.92**	NS	NS	NS	NS	NS	NS	NS	NS
$T_B(\alpha)$	NS	0.88**	NS	NS	NS	NS	0.95**	0.97**	NS	NS	NS	NS	NS	NS
$T_B(\alpha,\beta)$	NS	0.88**	NS	NS	NS	NS	0.95**	0.97**	NS	NS	NS	NS	NS	NS
$T_H(\alpha)$	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.83**	NS	NS	NS	NS
$T_H(\alpha,\beta)$	NS	NS	NS	NS	NS	NS	NS	NS	0.73*	0.88**	NS	NS	NS	NS
$T_V(\alpha)$	NS	NS	0.82**	NS	NS	NS	NS	NS	NS	NS	0.79*	0.83**	NS	NS
$T_V(\alpha,\beta)$	NS	NS	0.82**	NS	NS	NS	NS	NS	NS	NS	0.75*	0.79*	NS	NS
$T_F(\alpha)$	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	-0.71*	0.87**	NS
$T_F(\alpha,\beta)$	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.90**	0.70*

grid search. This is a much more plausible set of values, indicating that a source's emergent frequency components must be on average at most 5 dB less than other sources present at the same time for the source to be heard.

Table 1 shows the Pearson's correlation coefficients between the computed indicators and assessed parameters at the sound scene level (n=9). The three globally computed sound levels $L50$, $LA50$ and $L50_{1kHz}$ represent well the perceived overall loudness of the scene and can be used directly for pleasantness prediction. Ground truth emergences also correlate with the evaluated sound level parameters for all sources but traffic. The perceived time of presence is however represented poorly by both ground truth time of presence and emergence metrics: the source time of presence fails to account for potential masking by other sounds and the long-term emergence does not consider time distribution of activity. Furthermore, these indicators are not sufficiently discriminative of active sources as visible for traffic and birds or voices and footsteps. The $TFSD_{4kHz}$, while achieving good performances for the prediction of parameters related to bird presence, suffers from the same problem. The $TFSD_{500Hz}$ is not significantly correlated to the perception of human voices in this study. The two proposed emergence-based time of presence indicators achieve better performances: they are discriminative and show high correlations with the perception of corresponding sources. This confirms the need of an emergence-based presence indicator to successfully represent heard sources in the scene's mix.

For all sources the perceived time of presence and sound level are highly correlated ($r > 0.8, p < 0.01$). This is not the case for the corresponding acoustical indicators indicating information redundancy between these two quantities at the perceptual level. The sound level can thus be omitted in the pleasantness model. Furthermore, multiple linear regressions confirm that sources beside traffic, voices and birds have little impact on the studied perceptual notions of interest, liveliness and calmness. In these models similar contributions are obtained from other perceptual quantities and physical indicators.

4. DISCUSSION

We presented a pilot experiment to assess the relevance of predicting perceptual parameters from acoustic indicators in simulated scenes for soundscape quality assessment.

We find that the physical time of presence of sources is not sufficient to fully characterize soundscape perception. The proposed indicator $T_s(\alpha,\beta)$, while relying on a simple emergence model due to the small amount of available data, can be directly linked to source-specific perceptual quantities. It is more efficient for the assessment of traffic activity, which is heavily present in the background of sound scenes in a urban context. This illustrates the need to account for emergence of sources as a metric to determine their perceptual importance in complex sound scenes. Predicting the average pleasantness of a soundscape can thus be achieved by estimating the source activity and emergence indicators proposed in Section 2.

Precision requirements of the postulated physical metrics are also obtained. The 125 ms or longer time scales used for the computation of all indicators in the presented experiment are allow the design of perceptually relevant indicators. A binary (not heard - heard) masking model is shown in this study to improve parameter prediction. The estimation of source-wise emergence as a classification process (4 classes from *Not at all heard to Dominant*) as opposed to continuous regression is thus considered sufficient for the application needs.

Future work will consider 1) consider a refined perceptual experiment with a richer soundscape corpus in order to achieve a stronger validation and model design and 2) formulate a complete experimental protocol dedicated to the soundscape characterization task.

5. REFERENCES

[1] EC, "Directive 2002/49/ec of the european parliament and of the council of 25 june 2002 relating to the assessment and management of environmental noise," *Off. J. Eur. Communities*, vol. 189, p. 12, 2002.

[2] C. Mydlarz, J. Salamon, and J. Bello, "The implementation of low-cost urban acoustic monitoring devices," *Applied Acoustics*, vol. 117, pp. 207–218, 2017.

[3] F. Gontier, M. Lagrange, P. Aumond, A. Can, and C. Lavandier, "An efficient audio coding scheme for quantitative and qualitative large scale acoustic monitoring using the sensor grid approach," *Sensors*, vol. 17, 2017.

[4] J. Picault, A. Can, J. Ardouin, P. Crepeaux, T. Dhome, D. Ecotiere, M. Lagrange, C. Lavandier, V. Mallet, C. Mietlicki, and M. Paboeuf, "Characterization of urban sound environments using a comprehensive approach combining open data, measurements, and modeling," in *Acoustics '17, Boston*, 2017.

[5] R. M. Schafer, *The tuning of the World*, 1977.

[6] B. Berglund and M. Nilsson, "On a tool for measuring soundscape quality in urban residential areas," *Acta Acust. unit. Acust.*, vol. 92, pp. 938–944, 2006.

[7] A. Brown, "Towards standardization in soundscape preference assessment," *Applied Acoustics*, vol. 72, pp. 387–392, 2011.

[8] F. Aletta, J. Kang, and O. Axelsson, "Soundscape descriptors and a conceptual framework for developing predictive soundscape models," *Landsc. Urban Plan.*, vol. 149, pp. 65–74, 2016.

[9] O. Axelsson, M. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *J. Ac. Soc. Am.*, vol. 128, p. 2836, 2010.

[10] R. Cain, P. Jennings, and J. Poxon, "The development and application of the emotional dimensions of a soundscape," *Applied Acoustics*, vol. 74, pp. 232–239, 2013.

[11] B. D. Coensel and D. Botteldooren, "The quiet rural soundscape and how to characterize it," *Acta Acust. unit. Acust.*, vol. 92, pp. 887–897, 2006.

[12] P. Delaitre, C. Lavandier, C. Ribeiro, M. Quoy, E. D'Hondt, E. G. Boix, and K. Kambona, "Influence of loudness of noise events on perceived sound quality in urban context," in *Inter Noise*, 2014.

[13] P. Ricciardi, P. Delaitre, C. Lavandier, F. Torchia, and P. Aumond, "Sound quality indicators for urban places in paris cross-validated by milan data," *J. Ac. Soc. Am.*, vol. 138, pp. 2337–2348, 2014.

[14] P. Aumond, A. Can, B. D. Coensel, D. Botteldooren, C. Ribeiro, and C. Lavandier, "Modeling soundscape pleasantness using perceptive assessments and acoustic measurements along paths in urban context," *Acta Acust. unit. Acust.*, vol. 103, pp. 430–443, 2017.

[15] C. Lavandier and B. Defreville, "The contribution of sound source characteristics in the assessment of urban soundscapes," *Acta Acust. unit. Acust.*, vol. 92, pp. 912–921, 2006.

[16] M. Nilsson and B. Berglund, "Soundscape quality in suburban green areas and city parks," *Acta Acust. unit. Acust.*, vol. 92, pp. 903–911, 2006.

[17] A. Can, L. Leclerq, J. Lelong, and J. Defrance, "Capturing urban traffic noise dynamics through relevant descriptors," *Applied Acoustics*, vol. 69, pp. 1270–1280, 2008.

[18] A. Can, P. Aumond, S. Michel, B. D. Coensel, C. Ribeiro, D. Botteldooren, and C. Lavandier, "Comparison of noise indicators in an urban context," in *45th International Congress and Exposition of Noise Control Engineering*, 2014.

[19] L. Brocolini, C. Lavandier, M. Quoy, and C. Ribeiro, "Measurement of acoustic environments for urban soundscapes: choice of homogeneous periods, optimization of durations, and selection of indicators," *J. Ac. Soc. Am.*, vol. 134, pp. 813–821, 2013.

[20] M. Nilsson, D. Botteldooren, and B. D. Coensel, "Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas," in *19th International Congress on Acoustics*, 2007.

[21] J. Gloaguen, A. Can, M. Lagrange, and J. Petiot, "Creation of a corpus of realistic urban sound scenes with controlled acoustic properties," in *Proceedings of Meetings on Acoustics*, 2017.

[22] M. Rossignol, G. Lafay, M. Lagrange, and N. Misdariis, "Simscene : a web-based acoustic scenes simulator," in *1st Web Audio Conference (WAC)*, 2015.

[23] J. Jeon, J. Hong, C. Lavandier, J. Lafon, O. Axelsson, and M. Hurtig, "A cross-national comparison in assessment of urban park soundscapes in france, korea, and sweden through laboratory experiments," *Applied Acoustics*, vol. 133, pp. 107–117, 2018.

[24] M. Berzborn, R. Bomhardt, J. Klein, J. Richter, and M. Vorlander, "The ita-toolbox: an open source matlab toolbox for acoustic measurements and signal processing," in *43th Annual German Congress on Acoustics*, 2017.

[25] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.